Large Scale Science on NSF's Frontera System

John West, Paul A. Navrátil, Maytal Dahan, and Matthew Vaughn, *Texas Advanced Computing Center, University of Texas, Austin, TX, 78712, USA*

Planning for the Frontera supercomputer formally began in May 2017 when the U.S. National Science Foundation (NSF) invited proposals for a new leadership-class computing facility, the top tier of high-performance computing systems funded by the agency. The program was awarded to the Texas Advanced Computing Center (TACC) at the University of Texas at Austin in 2018, and Frontera began operations in 2019. The $60M system debuted at #5 on the November 2019 biannual TOP500 list of the world's largest systems.

Unlike more general-use supercomputers in the NSF's cyberinfrastructure portfolio, such as the Stampede2 supercomputer also hosted at TACC, which are designed to accommodate science workflows of all sizes, Frontera's primary mission is to help researchers solve very large-scale science problems. Where Stampede2 hosts more than 1,500 projects and nearly 10,000 users annually, Frontera's user base is roughly one hundred projects and hundreds of users, all of whom run at a very large scale.

Today Frontera remains the fastest supercomputer at any university, and one of the fastest in the world (#9 on the November 2020 TOP500 list)—a powerful, all-purpose tool for science and engineering. Frontera provides a uniquely balanced set of capabilities that support both capability and capacity simulation, data-intensive science, visualization, and data analysis, as well as emerging applications in AI and deep learning. Large supercomputing users find a familiar programming model and tools in a system that serves as a bridge to future exascale systems that will have many more cores and deeper memory hierarchies.

Frontera comprises a primary compute capability of nearly 41 petaFLOPS, consisting of more than 8,000 dual-socket servers with Intel 8280 ("Cascade Lake") processors in nodes provided by Dell. This computing capability represents a more than 5.5× increase over the Blue Waters base processors (Frontera was designated by NSF to replace Blue Waters at the National Center for Supercomputing Applications [NCSA], the previous NSF leadership system), and even greater memory capacity than the 22,000 nodes of that system. Nodes are connected via a Mellanox InfiniBand (IB) interconnect that provides very low latency, with 100 Gb/s to each node and 200 Gb/s between switches in a fat tree topology with minimal oversubscription. The complete system is housed in 121 racks and has a maximum system power of just over 5.8 MW.

The CPU-based primary system is augmented with additional GPU computing capabilities that address both single- and double-precision requirements, delivered via a mix of NVIDIA Quadro and Volta cards. Approximately, 2 TB/s of total storage bandwidth, 55 PB of usable Lustre disk-based storage, and 3 PB of all flash Lustre storage provide the capability to conduct next-generation science at unprecedented scales and broaden the system's relevance to the emerging data science community.

DATA-DRIVEN DESIGN

In designing the Frontera project, we found it instructive to examine workload data from Stampede1, a system in operation at TACC between 2012 and 2017 that debuted at #7 on the TOP500 list; the Blue Waters system Frontera was to replace; and the 54 US Department of Energy INCITE allocations awarded in 2017, which were divided equally between Oak Ridge National Lab's Titan and Argonne National Lab's Mira. Our study identified the most-used applications on Stampede1, Blue Waters, and INCITE. These were: partial differential equations (PDE) and molecular dynamics (MD), with lattice quantum chromodynamics (lattice QCD) and many-body/*n*-body problems representing important additional application categories. This analysis motivated specific decisions about the architecture most likely to effectively support users of Frontera.

PDE simulations have historically been large consumers of high-end computing cycles. We expect to

1521-9615 © 2021 IEEE
Digital Object Identifier 10.1109/MCSE.2021.3070962
Date of current version 15 June 2021.

continue over the operational life of Frontera. Many high-end PDE applications employ static-grid explicit solvers, which feature local interactions and, hence, scale well on highly parallel systems. Rarer than explicit solvers, but still important, are semiimplicit or fully implicit PDE solvers for which a global system solve is required, usually requiring sophisticated preconditioners to be effective. The challenge of a global system solves (and explicit methods that employ adaptive mesh refinement) is the complex and possibly dynamic memory access and communication patterns. But these challenges to large-scale parallelism are also being overcome. The 2015 Gordon Bell Prize-winning application (among many others) shows that, in principle, there is no reason why implicit and adaptive PDE solvers cannot scale with nearly ideal parallel efficiency to CPU-based systems with greater than million-way parallelism—provided sufficient effort is expended in algorithmic and software engineering. More challenging is obtaining good node performance, stemming from the low arithmetic intensity of typical stencil or sparse matrix operations associated with PDE solvers. But reasonable node performance can be achieved with requisite effort if local interactions are sufficiently dense and structured, which is often the case for highly nonlinear or multiphysics problems, or high-order discretizations. With their complex dynamic underlying algorithms and modest arithmetic intensity, and the considerable effort required to extract performance, GPU adoption was slow at the time and remains so. The most productive and performant option for most PDE solvers at design time was a CPU-based system combined with a fast, low-latency network.

The other large class of applications that dominate the usage of the largest HPC systems were MD, along with lattice gauge QCD and *n*-body/many-body problems, all of which can make excellent use of highly parallel distributed memory systems. MD, *n*-body, and many-body problems often feature physical phenomena that incorporate nonlocal interactions. In most cases, these can be coarse-grained, treated by fast Fourier transforms (FFTs), or neglected beyond some cutoff radius, and effective highly parallel solvers have been designed that exploit this feature. Dense localized interactions lead to high arithmetic intensity in these expensive kernels, meaning that high node performance is achievable. Lattice gauge QCD methods resemble PDEs in their execution of sparse, local matrix–vector products (but on a 4-D grid). Despite the challenges of parallelizing these applications at a large scale, good performance is now being achieved on high-throughput systems. In all of these application categories and similar to PDEs, a low-latency/high-bandwidth network is the key to scalability.

Overall, we concluded that a CPU-based primary system with powerful nodes and a fast network is the best choice to allow a broad spectrum of users with diverse needs to do their science at scale. For certain application classes that can make effective use of GPUs, such as MD and machine learning (ML), the single-precision GPU subsystem provides a cost-efficient path to high performance.

In addition to the large HPC projects identified through system usage analysis, it was important to consider emerging uses of advanced computing resources for which Frontera is now critical.

Prominent among these are data-driven and data-intensive applications, which the NSF Advisory Committee for Cyberinfrastructure (ACCI) recommended be acknowledged as a distinct discipline. There are two fundamental categories of data-driven science. In data assimilation and inverse problems, observational or experimental data are used to infer uncertain states or parameters in the (PDE, MD, etc.) model, which ultimately results in solution of the forward (PDE, MD, etc.) problem numerous times while intelligently exploring the parameter space, with added opportunity for task parallelism. In the second class of data-driven science, statistics, informatics, and analytics are used to learn directly from the data without recourse to the physics. These data applications often require nontraditional software (e.g., parallel R and MATLAB) and lead to the development of new data applications in which I/O read performance is even more important than write performance.

A nascent but growing class at the time of design (and well-established today) in this second category is ML. The coming decade will see significant efforts to integrate physics-driven and data-driven approaches to learning. We believed it was important that Frontera be designed with the capability to address very large problems in these emerging communities of computation, providing a well-balanced petascale HPC system with comprehensive capabilities that can serve a wide range of both simulation-based and data-driven science.

IN THIS ISSUE

Frontera was designed to apply to the widest possible breadth of science users, and the projects on the system range from astrophysics to COVID-19 modeling (a few Frontera users engaged in different aspects the COVID-19 pandemic response are featured in our November/December special issue). In this special issue, we include four articles from science users on the system and one written by a group at TACC

responsible for ensuring that both users and the system are ready for application runs that use all 8,000 nodes in a single job.

In "Modeling the World's Most Dangerous Thunderstorms," Leigh Orf *et al.* discuss the history and challenges of modeling the thunderstorms that spawn deadly EF4/5 tornadoes. Each year tornadoes are responsible for loss of life and property across large parts of the U.S. The storms that cause the most damaging tornadoes, supercell thunderstorms, have been the subject of numerical modeling since the dawn of the supercomputing age in the late 1970s. Orf highlights the role of computing in understanding these storm systems and the direct relationship between the quality of the forecast (and the amount of warning provided to communities in the path of these storms) and available computational capacity.

Stephen Yeager *et al.* review how future climate projections made with high-resolution models bridge the gap between weather and climate. In "Bringing the Future into Focus: Benefits and Challenges of High-Resolution Global Climate Change Simulations," we learn about recent simulations conducted by a team of climate scientists and software engineers that are yielding new insights into how climate change is driving weather extremes at regional scales.

David DeMarle *et al.* discuss developments in managing and analyzing the very large volumes of data that result from large-scale simulations of the type for which Frontera is deployed. They discuss a form of *in situ* visualization that permits the user to cache simulation data and go back in time to understand how a particular feature of interest evolved during the simulation.

"Towards improving the efficiency of organic solar cells by coarse-grained atomistic modeling of processing dependent morphologies," by Ganesh Balasubramanian *et al.* highlight the role that computing is playing in improving the efficiency of organic solar cells that harness the sun's energy for use as electricity. Balasubramanian reveals the role that computation plays in predictive design of new materials, using MD simulations to understand the process–structure–performance correlations in organic solar materials.

Finally, John Cazes *et al.* round out our discussion by giving us a view inside the science of running machine-scale jobs on a system the size of Frontera. "Preparing Frontera for Texascale Days" shows us how cyberinfrastructure providers and computational scientists work hand in hand with physical scientists to stage and complete runs that are often the largest ever attempted in their field. For Frontera, jobs at this scale are not only carried out during system benchmarks runs during commissioning or for Gordon Bell submissions—they are a routine part of Frontera operations and are providing scientists with a unique opportunity to push the boundaries of science.

The guest editors are especially grateful for the work of the reviewers, without whose detailed, thorough, and insightful comments this issue would certainly not be as polished and relevant as the version you are reading today. We want to make a special gesture to recognize and thank the reviewers for their all-too-often underappreciated contribution to the health of the academic community: Johannes Dahl (Texas Tech), Dan Dawson (Purdue University), Carina Farber (ENGIE), Christian Gribble (SURVICE Engineering Company), Pascal Grosset (Los Alamos National Laboratory), Patrick Heimbach (UT Austin), Dave Randall (Colorado State University), Josh Rhodes (UT Austin), Brett Roberts (Cooperative Institute for Mesoscale Meteorological Studies), Glen Romine (NCAR), Luis Paulo Santos (Universidade do Minho, Informatics), Joshua Schrier (Fordham University), and Phillip Wolfram (Los Alamos National Laboratory).

JOHN WEST is the Director of Strategic Initiatives with the Texas Advanced Computing Center (TACC), part of the University of Texas at Austin, Austin, TX, USA. TACC provides HPC services and expertise to the open science community and is the largest provider of HPC resources in the NSF XSEDE community. Prior to joining TACC, he was the Director of the USA Department of Defense High Performance Computing Modernization Program and held a number of positions in private industry and the federal government providing supercomputing resources and expertise for research and development missions. Contact him at john@tacc.utexas.edu.

PAUL A. NAVRÁTIL is currently a research scientist and Director of Visualization with the Texas Advanced Computing Center (TACC), University of Texas at Austin, Austin, TX, USA. He is an expert in high-performance visualization technologies, accelerator-based computing, and advanced rendering techniques. His research seeks to improve analytic capacity and insight communication across scientific workflows, including efficient algorithms for large-scale parallel visualization and data analysis (VDA), and innovative design for immersive VDA systems. His team provisions TACC's two visualization labs and the remote visual analytic environments on TACC's advanced computing systems, including the US NSF leadership-class systems *Stampede2* and *Frontera*. His work has been featured in numerous venues, both nationally

and internationally, including the *New York Times*, *Discover*, and *PBS News Hour*. He received the B.S., M.S., and Ph.D. degrees in computer science, and the B.A. degree in plan II interdisciplinary (hons.) from the University of Texas at Austin. Contact him at pnav@tacc.utexas.edu

MAYTAL DAHAN is currently the Director of Advanced Computing Interfaces (ACI) overseeing the Web and Mobile Applications and the Cloud and Interactive Computing groups, Texas Advanced Computing Center (TACC), part of the University of Texas at Austin, Austin, TX, USA. TACC provides HPC services and expertise to the open science community; The ACI team focuses on developing and deploying large-scale production-quality web, mobile and cloud computing projects, developing, deploying, and supporting user portals, science gateways, web and mobile interfaces, and APIs. Contact her at maytal@tacc.utexas.edu.

MATTHEW VAUGHN is the Director of Life Sciences Computing, Texas Advanced Computing Center, part of the University of Texas at Austin, Austin, TX, USA. He is a molecular biologist and technologist with more than 15 years' experience leading complex projects to design and deploy powerful collaborative platforms for research computing. Prior to joining TACC, he was a research assistant professor with the Cold Spring Harbor Laboratory, New York, where he investigated epigenetic natural diversity and the mechanisms of epigenetic gene silencing and coled creation of the CyVerse national life sciences cyberinfrastructure. Contact him at vaughn@tacc.utexas.edu.